

Class 12: Lab Session

Lana (PID: A17013518)

Section 1. Proportion of G/G in a population

Downloaded a CSV file from Ensembl < https://useast.ensembl.org/Homo_sapiens/Variation/Sample?db=core;v=rs8067378;vdb=variation;vf=959672880#373531_tablePanel >

Here we read this CSV file

```
mxl <- read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")
head(mxl)
```

| | Sample..Male.Female.Unknown.. | Genotype..forward.strand.. | Population.s. | Father |
|---|-------------------------------|----------------------------|---------------|--------|
| 1 | NA19648 (F) | A A | ALL, AMR, MXL | - |
| 2 | NA19649 (M) | G G | ALL, AMR, MXL | - |
| 3 | NA19651 (F) | A A | ALL, AMR, MXL | - |
| 4 | NA19652 (M) | G G | ALL, AMR, MXL | - |
| 5 | NA19654 (F) | G G | ALL, AMR, MXL | - |
| 6 | NA19655 (M) | A G | ALL, AMR, MXL | - |
| | Mother | | | |
| 1 | - | | | |
| 2 | - | | | |
| 3 | - | | | |
| 4 | - | | | |
| 5 | - | | | |
| 6 | - | | | |

```
table(mxl$Genotype..forward.strand.)
```

| | | | |
|-----|-----|-----|-----|
| A A | A G | G A | G G |
| 22 | 21 | 12 | 9 |

```
table(mx1$Genotype..forward.strand.) / nrow(mx1) * 100
```

| A A | A G | G A | G G |
|---------|---------|---------|---------|
| 34.3750 | 32.8125 | 18.7500 | 14.0625 |

Now let's look at a different population. I picked the GBR.

```
gbr <- read.csv("373522-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")
```

Find a proportion of G|G

```
round(table(gbr$Genotype..forward.strand.) / nrow(gbr) * 100, 2)
```

| A A | A G | G A | G G |
|-------|-------|-------|-------|
| 25.27 | 18.68 | 26.37 | 29.67 |

This variant that is associated with childhood asthma is more frequent in the GBR population than the MKL population.

Let's now dig into this further.

Section 4. Population Scale Analysis

Q13

One sample is obviously not enough to know what is happening in a population. You are interested in assessing genetic differences on a population scale. So, you processed about ~230 samples and did the normalization on a genome level. Now, you want to find whether there is any association of the 4 asthma-associated SNPs (rs8067378...) on ORM3 expression.

How many samples do we have?

```
expr <- read.table("rs8067378_ENSG00000172057.6.txt")
head(expr)
```

| | sample | geno | exp |
|---|---------|------|----------|
| 1 | HG00367 | A/G | 28.96038 |
| 2 | NA20768 | A/G | 20.24449 |
| 3 | HG00361 | A/A | 31.32628 |
| 4 | HG00135 | A/A | 34.11169 |
| 5 | NA18870 | G/G | 18.25141 |
| 6 | NA11993 | A/A | 32.89721 |

```
nrow(expr)
```

```
[1] 462
```

```
table(expr$geno)
```

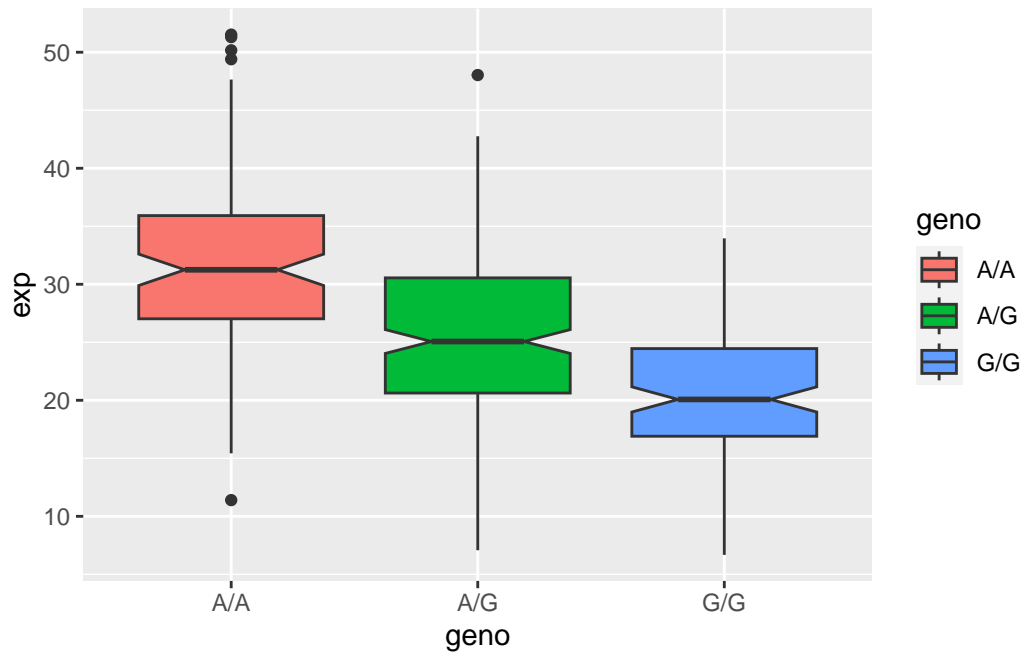
```
A/A A/G G/G
108 233 121
```

```
library(ggplot2)
```

Q14

Let's make a boxplot.

```
ggplot(expr) +
  aes(geno, exp, fill = geno) +
  geom_boxplot(notch = T)
```



The boxplot that I have made above shows that the SNP between G and A directly affects expression levels. The A|A median expression is ~32, the A|G median expression is ~ 25, and the G|G median expression is ~20. A|A has the highest expression of the three and based on the results of both A|A and A|G, the A allele most likely induces more gene expression.